SYNOPSIS

Ph.D in Computer Science

DEEP LEARNING – BASED FACIAL EXPRESSION RECOGNITION FOR ANALYZING LEARNER ENGAGEMENT IN MULSEMEDIA ENHANCED TEACHING

Submitted by

Ms. M.MOHANA

20PHCSF001

Supervisor

Dr. P. SUBASHINI

Professor

Department of Computer Science

Submitted to

Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore-641 043

August 2024

1. Introduction

The rapid growth and development of information and communication technologies (ICT) have enabled the fast evolution of Technology Enhanced Learning (TEL), a trend that is likely to continue. TEL refers to the use of technology to support and enhance the teaching and learning process, integrating various technological tools and platforms to improve educational experiences, making learning more accessible, engaging, and effective. The effectiveness and significant market potential of TEL has attracted increasing interest from individuals, corporations, and institutions. Notably, the global eLearning market is projected to reach a staggering $325 billion by 2025 [1].

1.1 Emotion Recognition in Learning

Emotion recognition in learning is an emerging technology designed to enhance the educational experience by identifying and responding to students' emotional states. Emotions play a crucial role in influencing motivation, engagement, and overall academic performance. In academic learning, emotions can be broadly categorized into positive and negative types: positive emotions like interest and joy can enhance learning, while negative emotions like frustration and anxiety can hinder it [2]. This technology can utilize various methods, including facial expression analysis, speech analysis, and physiological signals like heart rate variability. The goal is to improve the learning process by understanding learners' emotions and adapting responses accordingly.

Facial Expression Recognition in e-learning plays a significant role in enhancing the overall educational experience by focusing on understanding and responding to learners' emotions. According to research, the verbal component of a message, or spoken words, accounts for 7%, the vocal aspect (such as tone and pitch) makes up 38%, and facial expressions contribute 55% to the overall impact of the spoken message [3]. Ekman and Friesen proposed six basic emotions: happy, sad, surprise, fear, anger, and disgust. Sometimes researchers add neutral and contempt in this category [4]. These are commonly known as universal expressions or primary emotions. Facial emotions are identified through facial muscles moments as shown in Figure 1.

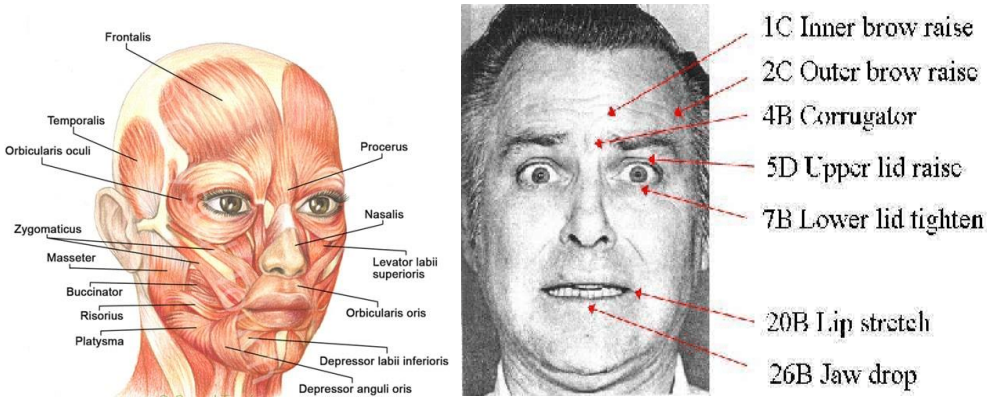


Figure 1. (a) Facial Muscles (b) Disgust facial expression muscle Movement converted into numerical code (Image Source: Google.com)

1.2 Mulsemedia in Learning

In recent years, interactive multimedia has enabled dynamic and immersive learning experiences by integrating various media such as audio, text, video, images, and animation. Unfortunately, this does not provide immersive learning experiences in real-time scenarios.

Therefore, mulsemedia, which stands for "Multiple Sensorial Media," refers to multimedia systems that engage multiple senses beyond just sight and sound. This technology aims to create more immersive and engaging experiences by incorporating various sensory stimuli such as taste, smell, touch, and even temperature into digital content [5].

Quality of Experience (QoE) [21] is used in the context of mulsemedia based learning to measure overall satisfaction and effectiveness of the learning experience from the learner's perspective. It encompasses various factors that contribute to how learners perceive, interact with, and benefit from the e-learning content and environment through collecting direct responses from users regarding their satisfaction. Furthermore, Mulsemedia has assessed two ways: one subjective (using a questionnaire) and the other objective (using physiological signals) [6][23].

1.3 Learner Engagement Analysis Using Facial Expression

Analyzing student engagement in TEL using facial expression recognition involves leveraging AI and computer vision to assess students' emotional states based on their facial expressions. Most facial expression recognition (FER) systems in learning environments are trained on universal facial expressions, which are then mapped into categories such as positive, neutral, and negative to analyze student engagement [11-13]. This technology captures real-time data through cameras and interprets these emotions, providing insights into how students are interacting with the material. By offering instant feedback, this method enables educators to adjust their teaching strategies, ultimately enhancing learning outcomes.

1.4 Problem Statement

Existing Facial Expression Recognition (FER) models often struggle to accurately analyze and interpret learners' engagement through universal expressions. Moreover, enhancing e-content with a multisensory experience is necessary to improve learners' Quality of Experience (QoE) by providing an immersive learning environment that enhances learning outcomes.

1.5 Research Objectives

1.5.1 Primary Objective

- To develop an effective FER model using a deep learning approach for predicting learners' engagement in mulsemedia-enhanced teaching.

1.5.2 Secondary Objective

- To experiment with various deep learning approaches (3D-CNN, LSTM, and various types of Autoencoder) for building effective FER systems in e-learning environments using a universal facial expression dataset.
- To minimize computational costs associated with FER to make the system more efficient and accessible.
- To maximize precision and recall scores to improve the overall reliability and effectiveness of the FER system.

2. Research Methodology

Figures 2 and 3 show the process of FER for Learner Engagement detection and the proposed approach for achieving aforementioned research problems.

Figure 2. Learner Engagement Detection Process in E-learning

Figure 3. Proposed Research Process Steps

2.1 Phase -1: Enhancing Viola-Jones face detection prediction accuracy by Particle Swarm Optimization (PSO)

Face detection is a computer vision task that involves identifying faces in images or video frames. It is a component of object detection and finds applications in various fields such as face recognition, facial expression analysis, security, biometrics, entertainment, and more. Achieving high detection accuracy in complex backgrounds is of paramount importance in real-time scenarios, where factors such as pose variations, varying illumination, occlusions, and scale variations pose significant hurdles.

a. Viola-Jones Algorithm

The Viola-Jones algorithm is a widely used framework for real-time object detection, especially for faces. It works by selecting key visual features using Haar-like patterns and quickly computing these features through an integral image technique. AdaBoost is then applied to identify the most relevant features, forming a strong classifier from many weaker ones. Finally, the algorithm uses a cascade of classifiers to efficiently scan the image, rapidly discarding non-object regions and focusing on areas likely to contain the target object. This combination allows for fast and accurate detection of images.

Limitation of Viola-Jones Algorithm

- Sensitive to variations in lighting conditions, which can affect the performance of the algorithm.
- Produce false negatives (missing faces that are present).
- Limit its ability to detect faces at different scales efficiently and provide multiple detection.
- Struggle when faced with occlusions, where faces are partially hidden.

Reasons identified for this Challenge in the Viola-Jones Algorithm after analyzing

- AdaBoost needs to search through all possible thresholds for all samples to find the minimum training error.
- Exhaustive feature search (1,16,000) consumes a significant amount of time to discover the best threshold values and optimize feature selection to build an efficient classifier for face detection.

2.1.1 Proposed Viola-Jones Face Detection with PSO Approach

Prediction accuracy and training error of the Viola-Jones algorithm have been reduced using the Particle Swarm Optimization (PSO) algorithm. The proposed solution to improve the prediction accuracy of the Viola-Jones algorithm on complex images is as follows:

- Firstly, Particle Swarm Optimization (PSO) is employed to dynamically select optimal threshold values for feature selection, thereby improving computational efficiency.
- Secondly, the feature selection process using AdaBoost within the Viola-Jones algorithm, integrating PSO to identify the most discriminative features for constructing a robust classifier.

2.2 Phase 2: Hybrid CNN-BiLSTM for Facial Expression Recognition using Sequence of images (Spatiotemporal Features)

Spatiotemporal features combine spatial and temporal information to provide a comprehensive representation of facial expressions. Spatiotemporal features integrate both static facial attributes (spatial) and their dynamic changes over time (temporal). This combination enables the recognition of complex facial expressions, micro-expressions (brief facial movements lasting a fraction of a second), and continuous emotional states.

Existing CNN-RNN architecture first uses a Convolutional Neural Network (CNN) to extract spatial features from the input data, typically images or video frames [14]. CNN learns to capture the spatial relationships and patterns within each frame or image. The output of the CNN is then fed into a Recurrent Neural Network (RNN) [18], such as an LSTM or GRU, which processes the sequence of CNN feature maps over time to model the temporal dynamics, which limits to the dynamic analysis of spatiotemporal features directly for analysis FER in real-time environments.

2.2.1 Proposed CNN-BiLSTM Method

The proposed CNN-BiLSTM architecture, as shown in Figure 4, utilizes Time Distributed layers instead of 2D CNN layers to directly learn both spatial and temporal features from the input data. Subsequently, LSTM (Long Short-Term Memory) networks are employed in a bidirectional manner to effectively identify dynamic changes in facial expressions. Parameters such as ELU activation [16], the Adam optimizer [20], a learning rate of 0.001, a batch size of 10, and batch normalization [19] have been used in this architecture.

This work begins with data collection from sources like the CK+ dataset [9] and custom datasets. Faces are detected using the Viola-Jones with PSO algorithm, after which the detected images undergo pre-processing steps including resizing, shuffling, and normalization. These pre-processed images are then stacked into frames to capture temporal information. The dataset is split into training, validation, and testing subsets, which are used to train a CNN-BiLSTM model—a combination that leverages a time-distributed layer for directly extract spatiotemporal features and

Bidirectional long short-term memory (BiLSTM) networks for analysis relation between previous inputs. After training, the model is evaluated based on accuracy, precision, recall, and F1-score metrics. The feedback loop indicates an iterative process, refining the deployment model based on evaluation results.

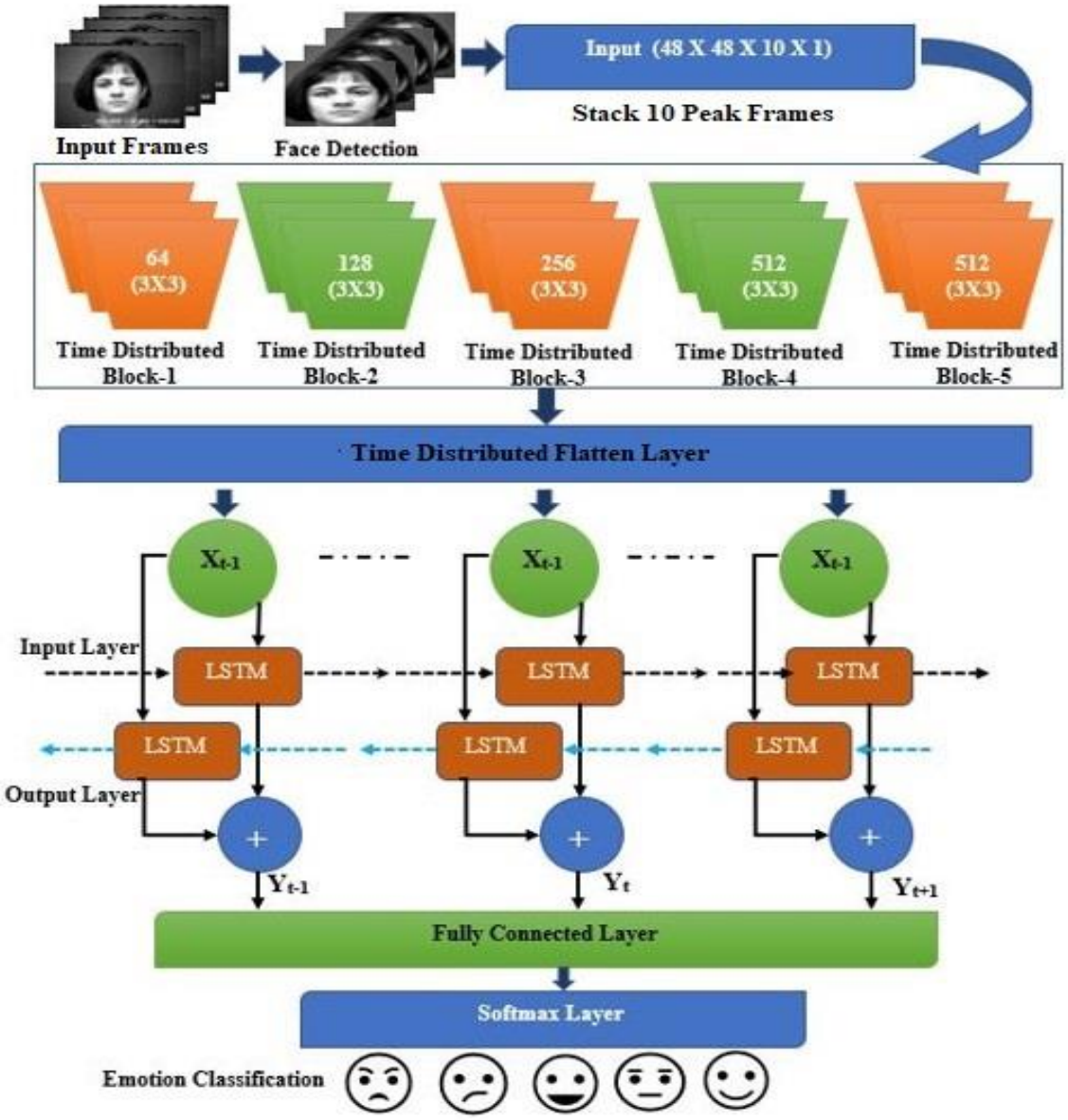


Figure 4. Proposed CNN-BiLSTM Architecture

2.3 Phase 3: Deep Semi-Supervised Convolutional Sparse Autoencoder

The supervised-based FER approach performance accuracy fully relies on a massive amount of labeled high-quality facial expression datasets. The problem with the existing FER datasets is that the total samples and expression counts on each emotion vary from dataset to dataset. This variation might be the reason for FER accuracy variations on different FER datasets. Moreover, increasing the layer size while training FER using CNN to create effective patterns for each expression on a small number of samples leads to higher computational costs [14-15]. However, this approach may result in overfitting issues and poor classification performance for some expressions in real-time environmental settings.

2.3.1 Proposed Approach

The newly introduced approach named as Deep Semi-supervised Convolutional Sparse Autoencoder to address the aforementioned issues and enhance FER performance and prediction accuracy. This approach comprises two parts: (i) Initially, a deep convolutional sparse autoencoder is trained with unlabeled samples of facial expressions. Here, KL (Kullback-Leibler) divergent sparsity is introduced in the convolutional block to enforce penalties, focusing on more relevant features for feature representation in the latent space. (ii) A trained encoder with a feature map is connected to a fully connected layer with softmax for final fine-tuning with learned weights and labeled facial expression samples in a semi-supervised approach for emotion classifications, shown in Figure 5.

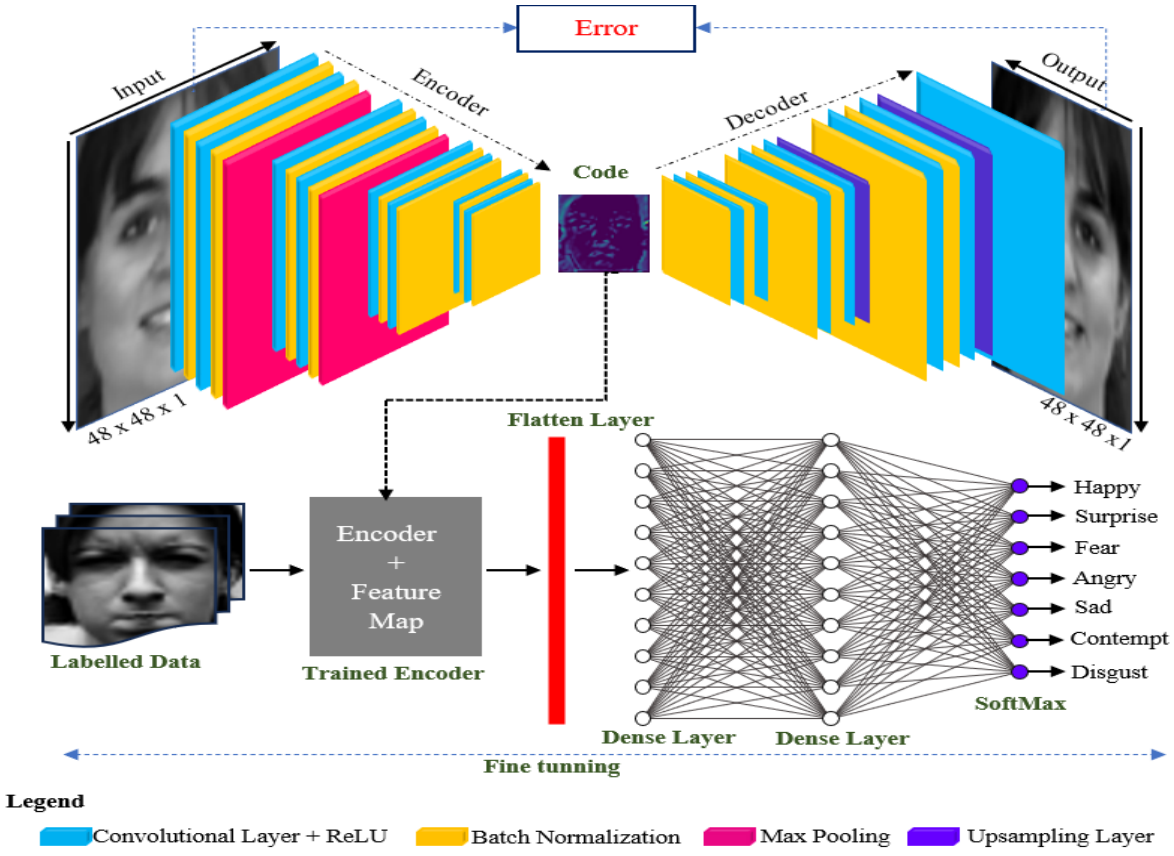


Figure 5. Deep Semi-Supervised Convolutional Sparse Autoencoder

2.4 Phase 4: Mulsemedia Delivery and analyse Learner's engagement through Facial Expression

Traditional multimedia applications in education usually focus on two human senses, audio and visual, and emphasize improving image and video quality. This limitation reduces interaction with digital content, separating users from computer-based multimedia. By incorporating a combination of audio and various other stimuli (e.g., olfactory, haptic, and gustatory), interaction with digital content can be increased. Currently, there is a growing focus on incorporating multisensory effects, including visual and audio information, haptic feedback, olfaction, and gustatory sensations, collectively known as Mulsemedia (Multisensorial media) (See Figure 6). This approach involves integrating multiple senses to create immersive and enriched learner experiences within the e-learning environment.

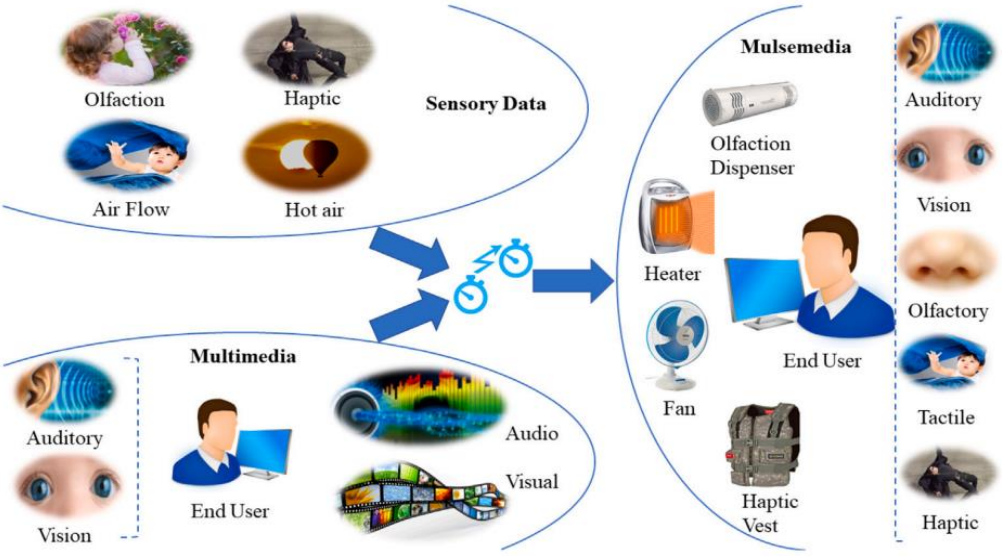


Figure 6. Mulsemedia Components, synchronization, and delivery to end users (Credit: google.com)

2.4.1 Designed E-Learning Portal with Mulsemedia Synchronization

There are two types of e-learning content has been created − specifically two videos, one targeting various biological aspects of the rosemary plant such as origin, appearance, smell, and their medical application and benefits, the other focusing on the physics behind the phenomena of thunder and lightning. The choice of the subject matter of the two videos was done with a view towards the rosemary video being enhanced with (rosemary) aromas for users in the experimental group, whilst the thunder and lightning video would be accompanied by airflow and vibrotactile effects, again for users in the experimental group. Both videos lasted 5 minutes and 28 seconds and were encoded and rendered at 30 frames per second as shown in Figure 7. The designed overall Mulsemedia components as shown in Figure 8.

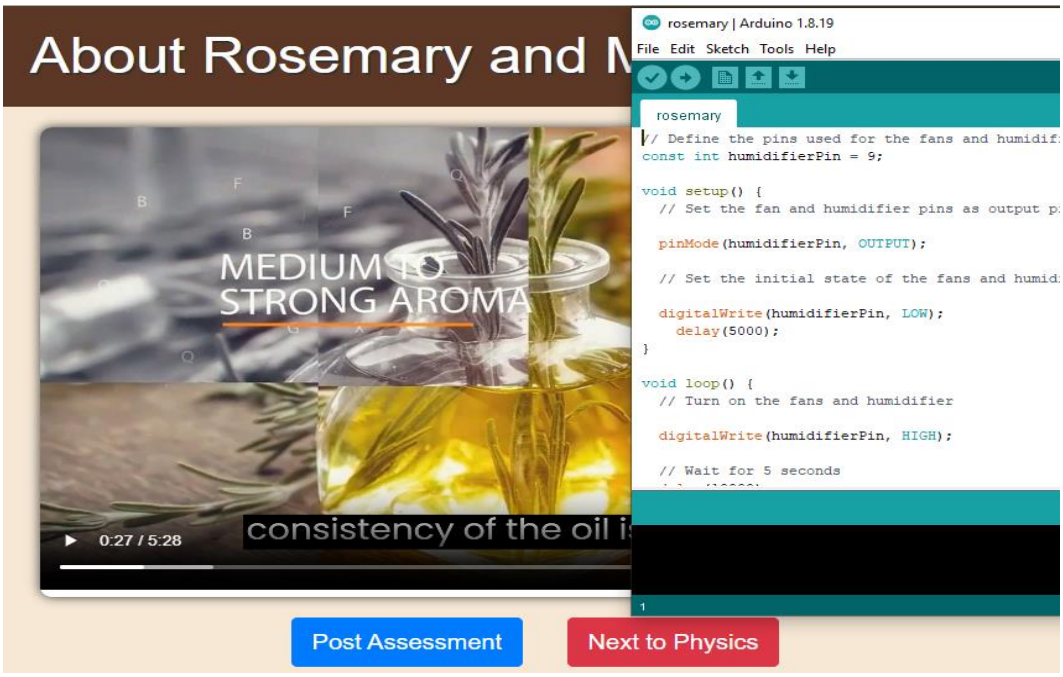

About Rosemary and M
MEDIUM TO
STRONG AROMA
consistency of the oil i
0:27 / 5:28
rosemary | Arduino 1.8.19
File Edit Sketch Tools Help
rosemary
// Define the pins used for the fans and humidif
const int humidifierPin = 9;
void setup() {
// Set the fan and humidifier pins as output p
pinMode(humidifierPin, OUTPUT);
// Set the initial state of the fans and humid
digitalWrite(humidifierPin, LOW);
delay(5000);
}
void loop() {
// Turn on the fans and humidifier
digitalWrite(humidifierPin, HIGH);
// Wait for 5 seconds
Post Assessment
Next to Physics

Figure 7. Video material featuring a rosemary plant to evoke the sensory effects of rosemary

Figure 8. Proposed Mulsemedia synchronized E-learning Portal (Olfactory, Haptic and Airflow)

2.4.2 Research Instruments

QoE Questionnaire: This consists of 22 questions designed to assess various aspects of the participant's experience during the e-learning process (See Table 1). The first 12 questions were only answered by experimental group participants, as they targeted the QoE in a mulsemedia context. Accordingly, the first four questions (Q1-Q4) focus on the impact of olfactory effects on the learning experience, including their enhancement of reality, the potential for distraction, and overall involvement. Q5 to Q8 explore the influence of airflow effects on the sense of reality, distraction, annoyance, and enjoyment in e-learning. Q9 to Q12 delve into the effects of vibration on the learning experience, including its enhancement of reality, potential for distraction, annoyance, and enjoyment. The remaining questions (Q13-Q22) evaluate the overall learning approach. They assess the ease of use, potential cumbersomeness, understanding of concepts, knowledge acquisition, satisfaction of learning needs, improvement of the learning experience, practical engagement in the learning process, enjoyment, recommendation to friends, and desire to continue learning with the experienced approach. Participants answered each QoE question on a 5-point Likert Scale ranging from Strongly Disagree to Strongly Agree, which for analysis purposes were coded with numerical values ranging from 1 to 5. The feedback gathered from these questions is used to shed light on the efficacy of the mulsemedia-enhanced e-learning system and its influence on learners' QoE.

Table 1 QoE Questionnaire

No.	Questions
Q1	The olfactory effects enhance the sense of reality in e-learning.
Q2	The olfactory effects are distracting.
Q3	The olfactory effects are annoying
Q4	I enjoy e-learning with olfactory effects.
Q5	The airflow effects enhance the sense of reality in e-learning
Q6	The airflow effects are distracting
Q7	The airflow effects are annoying
Q8	I enjoy e-learning with airflow effects
Q9	The vibration effects enhance the sense of reality in e-learning.
Q10	The vibration effects are distracting
Q11	The vibration effects are annoying

Q12	I enjoy e-learning with vibration effects
Q13	The experienced approach was easy to use
Q14	The experienced approach was cumbersome
Q15	The experienced approach helped me better understand the concepts explained
Q16	The experienced approach did not help me learn knowledge
Q17	The experienced approach can satisfy my learning needs
Q18	The experienced approach did not improve my learning experience
Q19	The experienced approach did not help me be more practically engaged in the learning process.
Q20	I enjoyed the experienced approach.
Q21	I would not recommend the experienced approach to my friends.
Q22	I would like to learn more with the experienced approach

3. Results and Analysis

3.1 Phase-1 Enhancing Viola-Jones face detection prediction accuracy by Particle Swarm Optimization (PSO)

3.1.1 Dataset Description

Face images were collected from the Yearbook Dataset of frontal-facing American high-school seniors [7], while non-face images were obtained from the Stanford Background Dataset [8]. These images are used for both training and testing purposes. These databases contain 4,999 different face images and 6,960 non-face images, all with a pixel resolution of 25x25. The positive and negative images are randomly divided into two folders. The training folder comprises 1,200 positive and 1,000 negative grayscale images (See Figure. 9)

Figure 9. Training and testing sample images (a) Face images (Positive Images) (b) non-face images (Negative Images)

3.1.2 Results and Analysis

The face detection algorithm has two classes: faces and non-faces. The performance of the proposed method is evaluated using Equations (1) and (2). The True Positive Rate (TPR) is used to measure how well the model correctly predicts the positive class. The equation for TPR is given below:

$$TPR = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)+False Negative (FN)}} \quad (1)$$

$$FPR = \frac{\text{False Positive (FP)}}{\text{False Positive (FP)+True Negative (TN)}} \quad (2)$$

3.1.3 Integration of Viola-Jones with PSO

The proposed approach is analyzed in terms of classifier accuracy and execution time. The experiments were repeated ten times for each algorithm, and the best, average, and worst outcomes are presented in Table 2. Additionally, the analysis indicates the number of features required for the face detection process on complex face images. The performance of face detection is influenced by both the population size and the number of PSO iterations. The results demonstrate the effectiveness of PSO for optimal feature selection in this problem when compared to the conventional Viola-Jones algorithm. In addition, Table 2 shows the number of features generated for a strong classifier during the training process. Viola-Jones with PSO required only 120 features in the best case, 134 in the average case, and 150 features in the worst case for building the weak classifiers, whereas the conventional Viola-Jones algorithm required 340 features in the best case. Therefore, the proposed method constructs superior classifiers using only 120 features in the best case, which is significantly fewer than the conventional Viola-Jones method.

Table 2. Number of features needed for detection

	Best	**Average**	**Worst**
Viola-Jones	340	340	340
Proposed	120	134	150

Table 3. Face Detection performance

Approach	**TPR (%)**	**FPR**	**Time**
Viola-Jones	96.63	0.0337	52.5s
Proposed	98.73	0.0127	30.6s

Table 3 summarizes the overall comparison between the conventional Viola-Jones and the proposed method on the ImageNet test dataset. The proposed approach achieved an average classification accuracy rate of 98.73% and an FPR of 1.27% on the face images when testing. In contrast, the conventional Viola-Jones algorithm achieved an average accuracy of 96.63% and a 3.37% FPR on the face images. These results indicate that the proposed method is not only more efficient than the Conventional-VJ algorithm but also more effective in classifying unseen datasets. Furthermore, the proposed method requires less time to test the dataset compared to the conventional algorithm.

3.2 Phase 2: Hybrid CNN-BiLSTM for Facial Expression Recognition using Sequence of images (Spatiotemporal Features)

3.2.1 Dataset Description

(a) In-house dataset comprises 40 female subjects aged 21 to 26 years, exhibiting variations in lighting, occlusion, and positions. The dataset includes a sequence of facial expressions from individuals who have given consent for research purposes. The experimental setup is depicted in Figure 10, and the entire process occurs in the Centre for Machine Learning and Intelligence (CMLI) laboratory. In addition, collected emotions are shown in Figure 11. A total of 1600 frames have been collected.

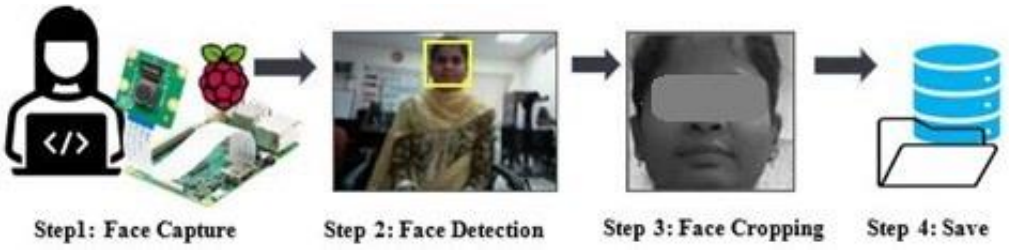


Figure 10. Framework for Facial Emotion Dataset Collection using IoT Kit

Figure 11. In-house datasets: happy (row 1), neutral (row 2), sleepy (row 3), surprise (row 4)

(b) The Extended Cohn-Kanade (CK+) [9] dataset consists of 593 image sequences featuring 123 different individuals. Among these images, 327 have been labeled to represent seven facial expressions: anger, happiness, sadness, surprise, fear, contempt, and disgust.

3.2.2 Experimental Setting

The FER datasets are divided into training (80%) and testing (20%) sets for this experiment. Performance data is obtained using the hold-out cross-validation method. The proposed network comprises nine TimeDistributed convolution layers with learning rates set at 0.001 and trained for 100 epochs. To address overfitting, an early stopping method is implemented. The evaluation involves CNN, LSTM, CNN with LSTM networks, and the proposed networks, conducted on Google Colab environments using Python and the Keras package with TensorFlow backends. Additionally, the experiments utilize graphics processing units (GPUs) installed in a Dell desktop featuring an Intel(R) Core (TM) i5-8400 CPU @ 2.80GHz 2.81 GHz.

3.2.3 Results and Analysis

Figures 12 and 13 illustrate the accuracy and cross-entropy (loss) performance assessments of the proposed model on the CK+ and In-House datasets throughout the training and testing phases. For the CK+ dataset, at epoch 100, the training and testing accuracies stand at 0.91% and 0.84%, respectively. Similarly, on the In-House dataset, the training and testing accuracies are 0.99% and 0.92% at epoch 50. Notably, the initial stage of training exhibits a slower convergence speed, followed by a more accelerated convergence around epochs 90–100 (refer to Figure 12 (a)), underscoring the model's satisfactory learning efficiency. The training curve demonstrates a reduced slope during model training. As shown in Figure. 13. (a), training accuracy gradually increases with slight fluctuations. Moreover, the CK+ and In-House datasets register training and testing losses of 0.84 and 0.61, respectively.

Figure 14 depicts the confusion matrix of the proposed model on both datasets. In Figure 14. (a), showcasing the CK+ dataset's performance, among 750 image sequences, 21 were misclassified due to facial appearance similarities across five emotions. Happy and surprised exhibited superior performance, benefiting from facial features and time series information, while anger, sadness, and fear were occasionally misclassified. For instance, anger was falsely classified as sad, sad as happy or angry, and fear as sad, angry, or happy. Precision, recall, and F1 scores for the four emotions ranged from 0.83 to 1.0 %, with an average precision and recall of 0.92 and 0.91%, respectively.

In Figure 14 (b), the confusion matrix for the In-House dataset illustrates that happy and sleepy achieved higher performance compared to neutral and surprised. Notably, neutral overlapped with sleepy expressions, and surprise slightly overlapped with neutral expressions. Precision, recall, and F1 scores for the five emotions varied from 0.38 to 1.0, with an average precision and recall of 0.85 and 0.84, respectively. Given the inherent challenges of facial expression recognition, including similarities in facial features and individual variations of the same expression, the proposed CNN with LSTM demonstrates notable results. It shows reliable true positive and true negative values, fewer false negatives and false positives, and consistent performance in categorizing emotion sequences.

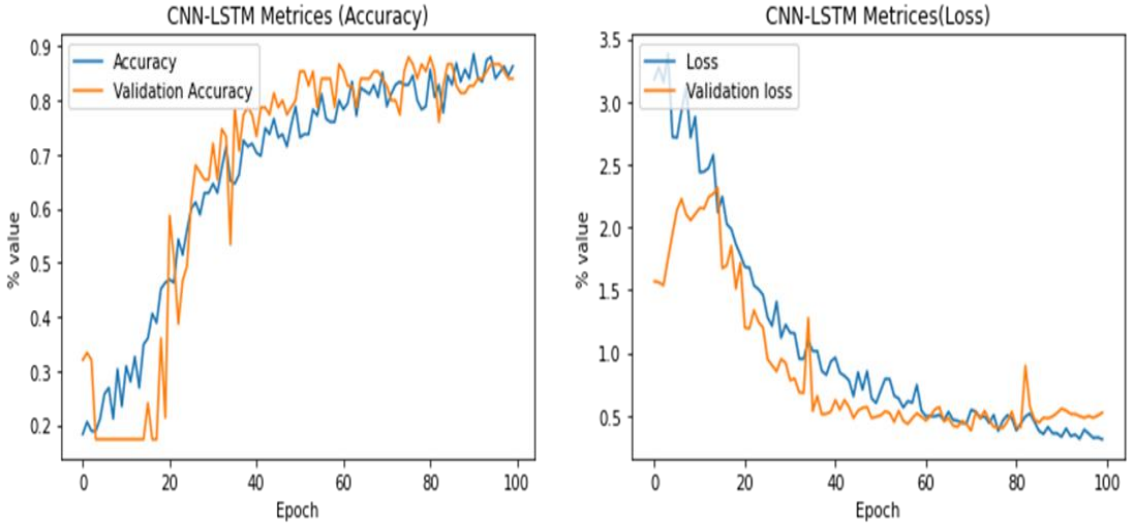


Figure 12. Timing process of CNN-LSTM in CK+ dataset (a) Accuracy (b) Loss

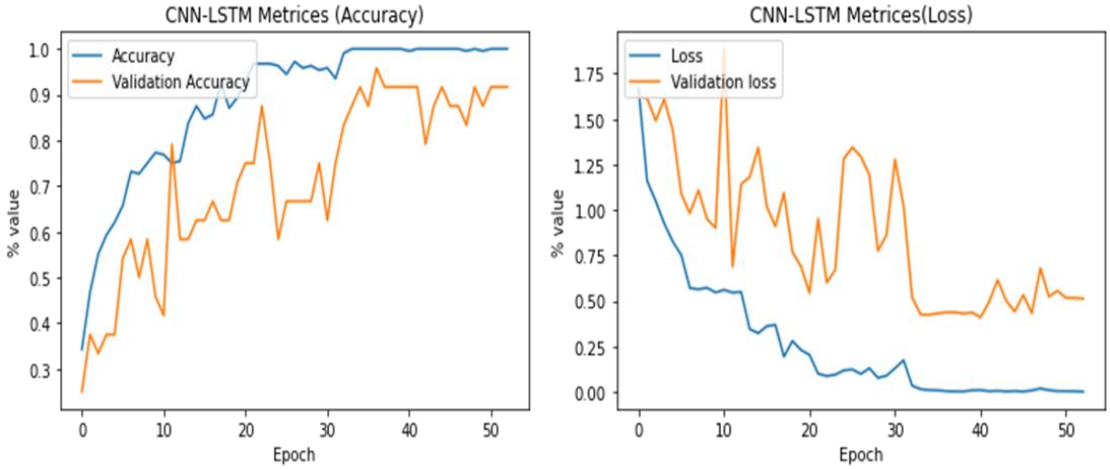


Figure 13. Timing process of CNN-LSTM in In-house dataset (a) Accuracy (b) Loss

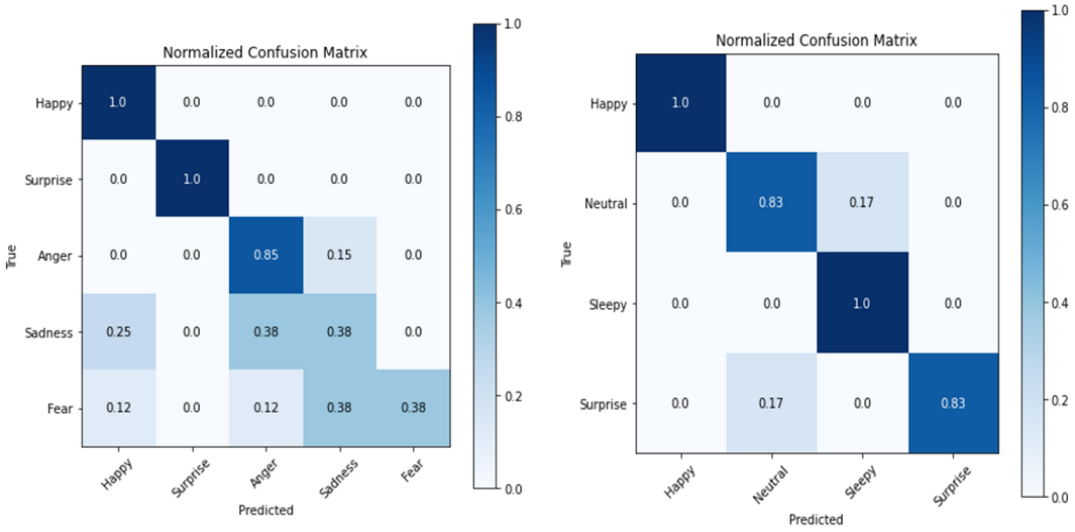


Figure 14. (a) Confusion matrix of CK+ (b) Confusion matrix of In-house dataset

Table 4. Parameter settings of each baseline model

Methods	Learning Rate	Epoch	Batch Size	Loss Function	Optimizers	Parameters
3D CNN	0.001	100	10	Cross entropy	Adam	76 M
LSTM	0.001	100	10	Cross entropy	Adam	57 M
CNN+LSTM	0.001	100	10	Cross entropy	Adam	49 M
Proposed Model	0.001	100	10	Cross entropy	Adam	46 M

Table 5. Comparison between different models on CK+ and In-house dataset

Methods	Accuracy CK+ (%)	Time	Accuracy In-house dataset (%)	Time
3D CNN	69.84	8 min	74.56	12 min
LSTM	50.78	10 min	79.38	16 min
CNN+LSTM	78.34	7 min	84.32	10 min
Proposed Model	**84.87**	**5.5 min**	**92.84**	**7.5 min**

Tables 4 and 5 show the results of the implementation of a CNN-BiLSTM method for extracting spatiotemporal features, enabling accurate emotion classification based on feature vector sequences. For this experiment, the in-house dataset was collected using a small 5MP camera attached to a Raspberry Pi, presenting challenges such as high illumination and noise. To mitigate these challenges, preprocessing techniques were employed before inputting data into the CNN-BiLSTM) model. Furthermore, the evaluation of benchmark datasets showcased the proposed model's performance, achieving 84.87% and 92.84% accuracy on the in-house and CK+ datasets respectively. Additionally, the efficiency of baseline models on the FER dataset was compared with the proposed CNN-BiLSTM), which effectively learns the sequential relationships in facial expressions to identify emotions.

Figure 15 shows Grad-CAM [10] visualization of the CNN-BiLSTM model's prediction results from the last CNN layer with correctly identified labels. Grad-CAM (Gradient-weighted Class Activation Map) is a deep learning technique used to visualize and understand the decisions made by the CNN model. Grad-CAM computes the gradient of the predicted class score concerning the feature maps of the last convolutional layer, determining the importance of each feature map for a

specific class by analyzing these gradients. In Figure 15, the CK+ dataset on the top depicts the most highly concentrated features on facial expressions for emotion classification, such as the raised lip corners contributing to the classification of happy emotions, nose wrinkles contributing to disgusted emotions, and open mouth with raised upper lids for surprise expressions. Similarly, the proposed model has been shown to predict sleepy expressions with closed eyelids using the in-house dataset.

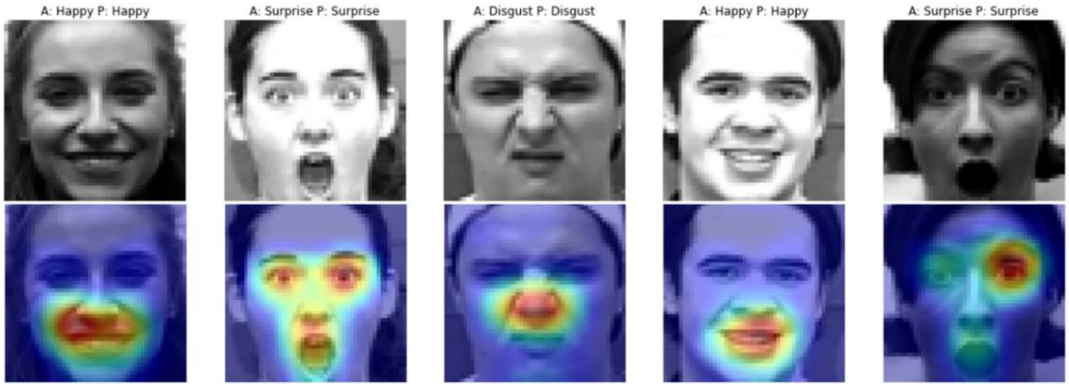


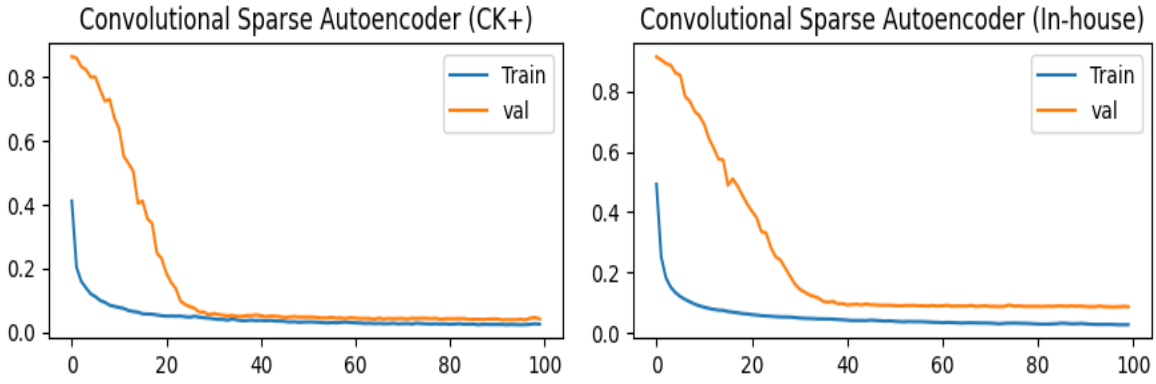
Figure 15. Grad-CAM Visualization of correctly Identified FER samples

3.3 Phase 3: Deep Semi-Supervised Convolutional Sparse Autoencoder

3.3.1 Results and Analysis

For this experiment, the DSCSA architecture has been evaluated using the following metrics: TP for True Positive, FN for False Negative, TN for True Negative, and FP for False Positive. In addition, The Mean Squared Error (MSE) is utilized to compute the autoencoder's reconstruction loss, while the Peak Signal-to-Noise Ratio (PSNR) is employed to compare the quality of the original and compressed reconstructed images.

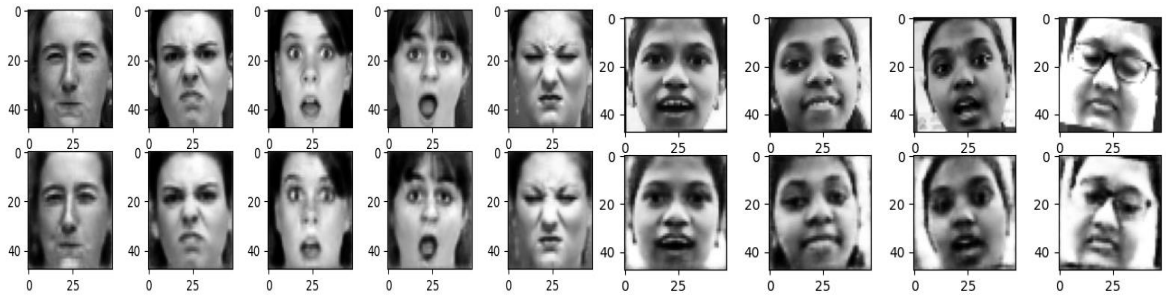
The proposed model underwent training for 100 epochs with a batch size of 32. These hyperparameter values were selected after several tuning iterations to achieve better results and minimize the loss values. Figure 16 shows the loss of deep convolutional autoencoder.



(a) (b)

Figure 16. Loss graph of DSCSA of CK+ and In-house

In Figure 17(a), the reconstruction results of the CK+ dataset for different expressions are illustrated, while Figure 17(b) displays the results for the In-house dataset. The obtained values indicate that the proposed autoencoder has effectively preserved essential features and minimized errors in facial expressions when reconstructing the original images as output images.



(a) (b)

Figure 17. First row: Original Image Second row: Reconstructed Image

Tables 6 and 7 display the precision, recall, F1-score, accuracy, training, and testing times for each model's performance on the CK+ and In-house datasets. The results demonstrate the effectiveness of the proposed model across different baseline models, despite the imbalanced nature of the dataset and the lower number of samples for each expression, particularly achieving notable performance on both CK+ and In-house datasets. Nevertheless, the baseline models did not perform well and failed to yield significant results under these conditions, likely due to challenges in identifying significant facial feature patterns with a limited number of data samples. Furthermore, the proposed autoencoder with sparsity helped to ignore irrelevant features and focus on efficient features during decoding, resulting in improved accuracy compared to the baseline AE (Autoencoder) performance.

Table 6. Experimental and comparison Results on CK+

Performance	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**Training Time**	**Testing Time**	**PSNR (dB)**
CNN	0.67	0.63	0.63	0.77	80s	6s	-
DenseNet 121	0.95	0.94	0.94	0.94	74s	4s	-
AE	0.95	0.96	0.96	0.96	60s	2s	68.65
DSCSA	**0.97**	**0.99**	**0.98**	**0.98**	**55s**	**2s**	**70.67**

Table 7. Experimental and comparison Results on In-house

Performance	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**Training Time**	**Testing Time**	**PSNR (dB)**
CNN	0.66	0.67	0.66	0.66	84s	4s	-
DenseNet 121	0.72	0.73	0.73	0.73	72s	3.2s	-
AE	0.83	0.84	0.83	0.83	60s	3s	64.09
DSCSA	**0.98**	**0.98**	**0.98**	**0.98**	**53s**	**2.8s**	**67.04**

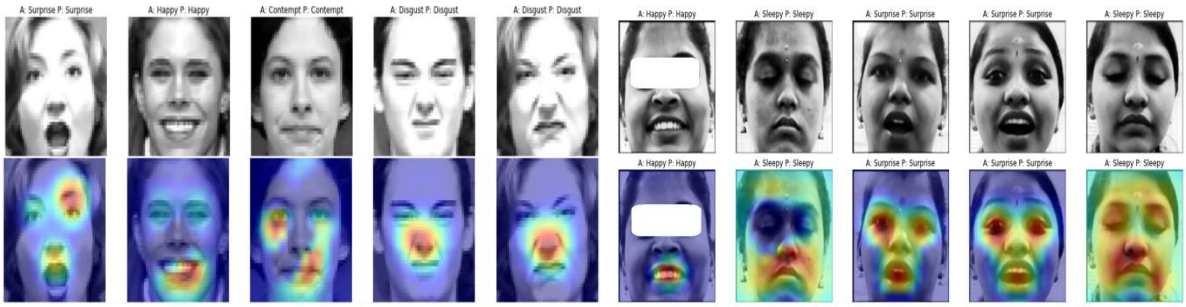


Figure 18. Grad-CAM Visualization of CK+ and In-house

Figure 18 displays the Grad-CAM visualization of CK+ and JAFFE samples with various facial expressions, clearly illustrating how the model accurately predicted emotion labels and presented

features contributing to that decision. Noteworthy features, such as an open mouth during surprise, pulled mouth corners when smiling, tightened lips during contempt, and nose wrinkles during disgust, are significant for identifying these expressions. These features are also described in Facial Action Units (AU) for classifying facial expressions.

3.4 Phase 4: Mulsemedia Delivery and Analysis of Learner's Engagement through Facial Expression

3.4.1 Mulsemedia Delivery Assessment (Pilot Study)

3.4.1.1 Participants

The research study was conducted at the Centre for Machine Learning and Intelligence Laboratory, Avinashilingam University, India, with a group of 70 (35 control group, 35 experimental) participants, all aged between 21 and 30 years. About 40% of these participants were undergraduates in the age group of 21-23, another 40% were postgraduates aged between 23-25, and the remaining 20% were researchers aged between 25-30, pursuing STEM degrees.

Each participant was randomly assigned to one of two equal-sized experimental and control groups. Experimental group (EG) learners were exposed to e-content enhanced with mulsemedia (incorporating olfactory, airflow, or vibrotactile effects). In contrast, control group (CG) learners experienced the same content.

3.4.1.2 Results and Analysis

Comparative analysis of learning performance between the Control Group (CG) and the Experimental Group (EG) reveals some interesting insights. A 10-question multiple-choice questionnaire was created for both versions of the e-content to assess learners' recall of information, with and without the effects of multisensory media. Accordingly, as shown in Table 8 for both CG and EG, there was an increase in the mean learning performance from the pre-test to the post-test, irrespective of the visualized content (Ros or TL). Moreover, a paired samples t-test revealed that, whilst there were no statistically significant differences in pre-test scores between CG and EG participants for either biology or physics content (Pair 1 and Pair 3), the learning performance for EG participants in post-tests was higher – and statistically significant − than that of their CG counterparts, for both types of subject matter (Pair 2 and Pair 4). Thus, in this study, exposure to mulsemedia has resulted in statistically significant higher learning performance.

Table 8. Paired Sample T-Test

		Mean	t	df	Sig. (2-tailed)
Pair 1	CG_Ros_Pre - EG_Ros_Pre	3.429	1.046	34	.303
Pair 2	CG_Ros_post - EG_Ros_Pos	-20.286	-7.875	34	.000
Pair 3	CG_TL_Pre -EG_TL_Pre	2.857	.935	34	.356
Pair 4	CG_TL_Pos - EG_TL_Pos	-25.143	-7.209	34	.000

The comparison between the CG and the EG indicates substantial differences in how participants perceived and experienced the educational approach (See Table 9). The EG consistently rated the approach more positively across various measures. Participants in the EG found the approach notably easier to use (4.86) compared to the CG (3.94), indicating a higher level of user-friendliness. Additionally, the EG perceived the approach as significantly less cumbersome (1.51) in contrast to the CG (3.20), suggesting a stark contrast in perceived difficulty levels. In summary, the EG consistently rated the experienced approach more positively across various dimensions

compared to the CG. These findings suggest that the intervention or the method applied to the EG had a more favorable impact on user experience, perceived learning outcomes, engagement, and motivation compared to the CG.

Table 9. Comparison of The Control Group and Experimental Group

Items	Group	Mean (Std. Dev.)	P-value (t-value)
Experienced approach was easy to use	CG	3.94 (.482)	.000
	EG	4.86 (.355)	(-9.040)
Experienced approach was cumbersome	CG	3.20 (.797)	.000
	EG	1.51 (.612)	(9.923)
Helped understand the concepts	CG	3.91 (.507)	.000
	EG	4.86 (.355)	(-9.011)
Did not help learn knowledge	CG	2.17 (.707)	.000
	EG	1.29 (.622)	(5.568)
Can satisfy learning Needs	CG	3.71 (.710)	.000
	EG	4.86 (.355)	(-8.517)
Did not improve learning Experience	CG	2.40 (.812)	.000
	EG	1.23 (.426)	(7.560)
Did not help to be practically Engaged	CG	3.37 (.910)	.000
	EG	1.17 (.382)	(13.183)
Enjoyed the experienced approach	CG	3.71 (.622)	.000
	EG	4.91 (.284)	(-10.386)
Would not recommend to friends	CG	2.57 (.850)	.000
	EG	1.17 (.382)	(8.885)
Would like to learn more	CG	3.66 (.684)	.000
	EG	4.89 (.323)	(-9.615)

3.4.2 Learner Engagement Analysis in enhanced Mulsemedia teaching

3.4.2.1 Participants

For this assessment, 20 participants were included, divided equally between two groups: an experimental group and a control group. Each group consisted of 10 participants. All participants were postgraduate or PhD students with science backgrounds.

Initially, consent forms were provided to participants for this study. Ethical clearance was obtained from the Institutional Human Ethics Committee of Avinashilingam University for Mulsemedia in Affective Computing (Approval number: AUW/IHEC/CS-22-23/XPD-07). As mentioned, the study involved two groups. The experimental group received Mulsemedia-synchronized content, with participants' facial expressions and predicted emotions recorded. At the

same time, predicted learners' engagement results were recorded individually and stored in Excel for overall analysis. The control group followed the same procedure without Mulsemedia effects. The overall architecture of the learner's engagement analysis has been provided in Figure 19.

3.4.2.2 Procedure

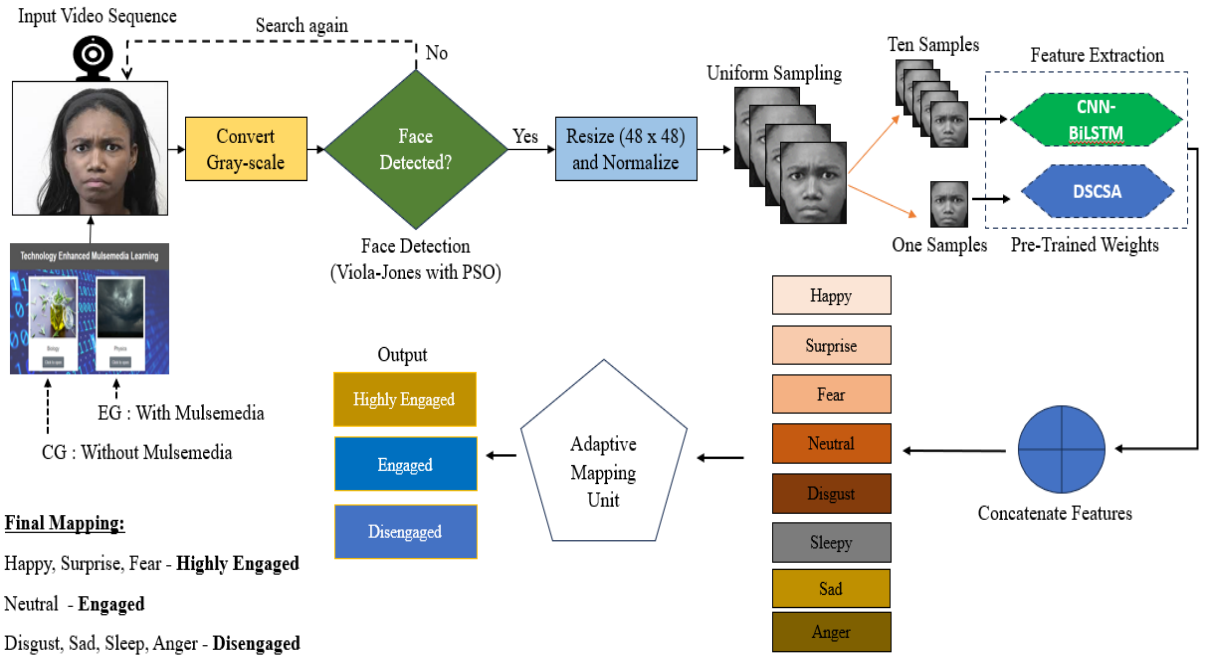


Figure 19. Overall architecture e-learning emotion assessment

Figure 19. illustrates a process designed to assess learners' engagement through facial expression analysis using advanced deep-learning techniques. The system begins by processing an input video sequence, converting it to grayscale to simplify further analysis. It then detects faces using a Viola-Jones algorithm enhanced with Particle Swarm Optimization (PSO). Upon successful face detection, the face is resized to 48x48 pixels and normalized. Multiple frames are sampled uniformly, with ten samples being used to extract both spatial and temporal features through a CNN-BiLSTM architecture, and a single sample processed using a DSCSA network with pre-trained weights. The extracted features are concatenated and passed to an adaptive mapping unit, which classifies the expressions into emotions such as happiness, surprise, fear, neutrality, disgust, sleepiness, sadness, and anger. These emotions are further mapped to engagement levels: "Highly Engaged" (happy, surprised, fear), "Engaged" (neutral), and "Disengaged" (disgust, sad, sleepy, anger). The system is designed to differentiate between groups receiving multisensory-enhanced content (EG) and those without it (CG), ultimately aiming to improve the learning experience by understanding and responding to learner engagement.

3.4.2.3 Learner Engagement Detection from Facial Expression

The system utilizes Mulsemedia synchronized e-content to evaluate engagement. A web camera captures the real-time facial emotions of learners, which are then used to determine the Engagement Index (EI). This index categorizes engagement into three states: highly engaged, engaged, or disengaged, and also provides an engagement level in percentage terms. The system recognizes a total of eight facial emotion classes, trained using three different datasets: CK+, JAFFE, and an In-house dataset. These classes are categorized into the three engagement states, as

detailed in Table 10 The predicted emotion is used as input to determine the engagement state. The Engagement Index (EI) is calculated using Equation (1), based on the predicted emotion's value to determine the engagement states.

$$EI = \sum(EP_i \, X \, WE_i) \quad (1)$$

- where EP = Emotion Probability for emotion i (Emotion = Neutral, Angry, Sad, Happy, Surprised, Sleepy, Disgust, and Fear)
- WE = Weight of the corresponding Emotion i

The EP (emotion probability) score is generated by a fusion classifier consisting of CNN-BiLSTM and DSCSA models, along with the corresponding emotion weight. Emotion Weight describes the emotional state's value that reflects the learner's engagement at that moment. The engagement percentage is then calculated based on the Engagement Index given in Equ (1). The weights for corresponding emotions are scaled and represented in Table 11. The algorithm for calculating engagement based on facial emotion recognition is provided below.

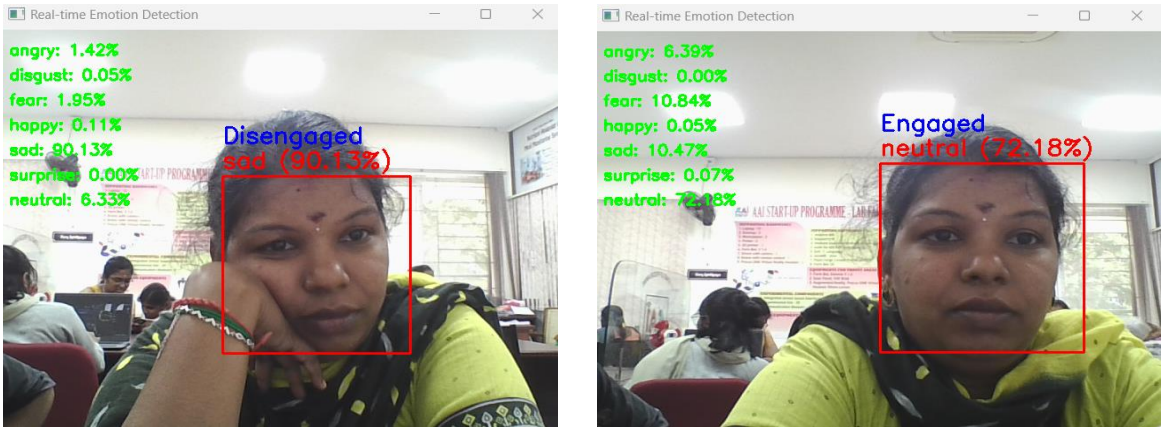
Table 10. Learners Engagement Classification

Emotion	Engagement
Happy (EP ≥ 50)	Highly Engaged
Surprise (EP≥ 80)	Highly Engaged
Neutral (EP ≥ 80)	Highly Engaged
Happy (EP < 50)	Engaged
Surprise (EP > 60 and EP < 80)	Engaged
Neutral (EP > 60 and EP < 80)	Engaged
Neutral (EP > 60) + Angry (EP <20) + Sad (EP > 30)	Engaged
Angry (EP > 20)	Disengaged
Sad, Sleep	Disengaged
Fear (EP > 30)	Disengaged
Angry (EP > 20) + Sad (EP > 30)	Disengaged
Angry (EP > 20) + Sad (EP > 30) + Fear (EP > 30)	Disengaged
Sad (EP > 30) + Fear (EP > 30)	Disengaged
Fear (EP > 30) + Surprise (EP < 60)	Disengaged
Neutral (EP < 60)	Disengaged

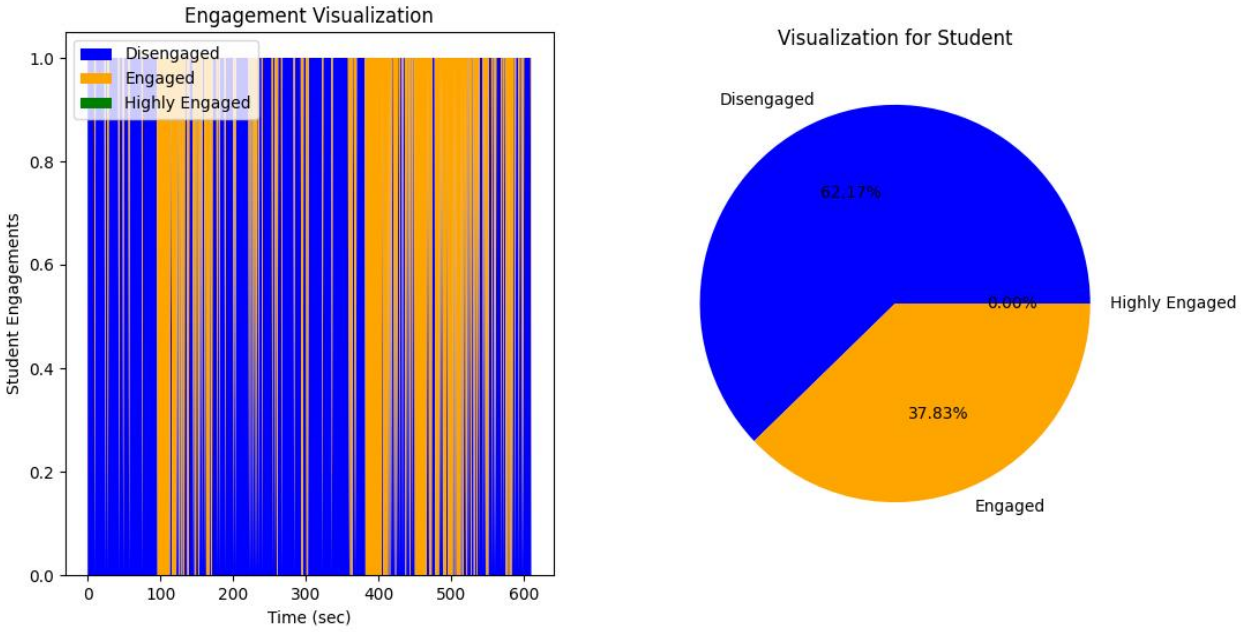
Table 11. Weight for the corresponding Label

Emotion	Weight Value
Happy	0.8
Suprise	0.8
Fear	0.6
Neutral	0.5
Angry	0.3
Sad	0.3
Sleepy	0.2
Disgust	0.3

Figure 20. shows that learners without mulsemedia effects felt somewhat bored while watching the e-content, which may explain the fluctuations between engagement and disengagement. Ultimately, the results indicated that learners were more disengaged than engaged with the content.



(a) (b)



(c)

Figure 20. (a) and (b) show participant engagement and disengagement detection without mulsemedia effects. (c) presents the overall results of learner engagement without mulsemedia.

Figure 20 shows two visualizations related to student engagement. The left graph, titled "Engagement Visualization," depicts engagement levels over time for individual students, using a color-coded scheme where blue represents disengaged, orange represents engaged, and green represents highly engaged students. The x-axis represents time in seconds, while the y-axis represents student engagement as a proportion. This graph shows frequent shifts between different engagement levels throughout the observed period. The right pie chart, titled "Visualization for Student," illustrates the overall engagement distribution, revealing that 62.17% of students are disengaged, 37.83% are engaged, and none are highly engaged. This suggests a predominance of disengagement among students with some engagement but no highly engaged students when multimedia learning. In addition, During the assessment of individuals' emotional reactions and their likelihood of being highly engaged, engaged, or disengaged, the data was stored in a CSV (Comma-Separated Values) file, as shown in Figure 21, for the final evaluation.

Control Group

Participants	Time Step	Dominant Emotion	Emotion Percentage	Engagement
1	0	sad	66.53683186	Disengaged
	1	sad	80.87893724	Disengaged
	2	sad	49.4366467	Disengaged
	3	sad	91.76254819	Disengaged
	6	sad	68.43494177	Disengaged
	7	sad	58.52235657	Disengaged
	8	sad	43.22674057	Disengaged
	9	neutral	47.66984852	Disengaged
	10	neutral	67.3910737	Engaged
	11	neutral	48.61036539	Disengaged
	12	angry	37.92746663	Disengaged
	13	neutral	40.63746929	Disengaged
	14	sad	37.72041426	Disengaged
	15	neutral	46.78665698	Disengaged
	16	neutral	38.5661602	Disengaged
	17	neutral	37.64521778	Disengaged
	18	sad	38.41333251	Disengaged
	19	angry	39.72476125	Disengaged
	20	sad	38.31894835	Disengaged
	21	sad	38.30789854	Disengaged
	22	sad	67.20539537	Disengaged
	23	neutral	38.08687627	Disengaged
	24	neutral	57.39126205	Engaged
	25	sad	58.29239239	Disengaged

Figure 21. Emotion probability and engagement detection results were recorded during the assessment without mulsemedia effects.

In a similar way, 10 participants were involved in the control group, and another 10 participants were involved in the experimental group. Each participant's emotional probability has been stored individually as a CSV file. Afterward, the results were combined and analyzed to find the overall difference in satisfaction and engagement levels between the two groups in the e-learning environment. The combined results are shown in Figure 22.

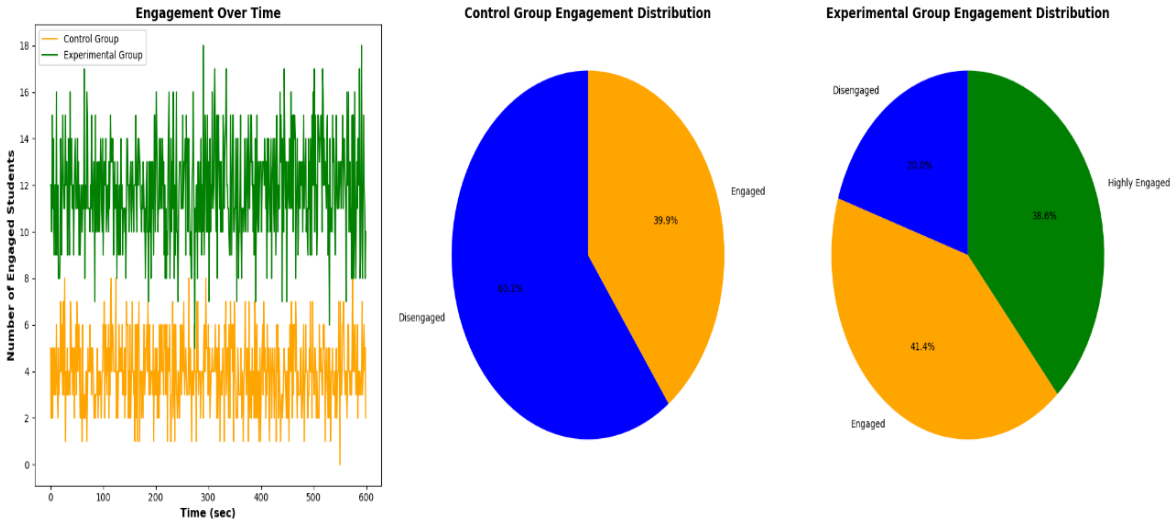


Figure 22. Plotting the overall Real-time Engagement Index of both Group Results

In summary, Figure 22 presents data on student engagement in both control and experimental groups through three different visualizations. The left plot shows the number of engaged students over time, where the control group (green) maintains a higher and relatively stable engagement level compared to the experimental group (orange), which shows lower engagement throughout the observed period. The middle pie chart illustrates the engagement distribution in the control group, with 60.1% of students being disengaged and 39.9% engaged. The right pie chart

displays the engagement distribution in the experimental group, where 20.0% are disengaged, 41.4% are engaged, and 38.6% are highly engaged. These charts highlight a more varied engagement outcome in the experimental group compared to the control group.

4. Conclusion

E-learning has grown significantly, offering flexible education opportunities. Emotion recognition (ER) enhances this by understanding students' emotions, leading to personalized learning and better engagement. However, challenges exist in current Facial Emotion Recognition (FER) systems, such as detecting faces under varied conditions and the need for large datasets in supervised learning.

4.1 Summary of overall contribution:

- Enhanced Viola-Jones Algorithm: Integrating Particle Swarm Optimization (PSO) improves face detection accuracy by optimizing feature selection and threshold values, boosting prediction accuracy by 2.1%.
- Hybrid CNN-BiLSTM Model for FER: Combines spatial and temporal features for real-time emotion detection, achieving high accuracy on benchmark datasets.
- Semi-supervised Convolutional Sparse Autoencoder: Reduces reliance on labeled data, offering high accuracy and interpretability in emotion recognition.
- Mulsemedia-Enabled Learning: Introduces multisensory experiences in e-learning, enhancing engagement and learning outcomes compared to traditional methods.

4.2 Limitation

- Focusing only on female facial expressions in data collection to train a deep learning model may lead to biased results when including other genders and may have limitations in real-world applications.
- Facial expressions alone do not reveal the true emotions of learners, so there is a need to involve physiological signals for more accurate emotion recognition.
- Cultural and age differences may influence the performance of facial expression prediction when training models with specific cultural or age groups.

References

1. Marín, V. I., de Benito, B., & Darder, A. (2024). Technology-enhanced learning for student agency in higher education: A systematic literature review. *Interaction Design and Architecture (s) Journal-IxD&A, 2020, num. 45, p. 15-49*.
2. Imani, M., & Montazer, G. A. (2019). A survey of emotion recognition methods with emphasis on E-Learning environments. *Journal of Network and Computer Applications*, *147*, 102423.
3. Begum, F., Neelima, A., & Valan, J. A. (2023). Emotion recognition system for E-learning environment based on facial expressions. *Soft Computing*, *27*(22), 17257-17265.
4. Ekman, P. (1972). Universals and Cultural Differences in Facial Expressions of Emotions. In Cole, J. (Ed.), Nebraska Symposium on Motivation (pp. 207-282). Lincoln, NB: University of Nebraska Press.
5. Orawetz, J., & Mühlberg, F. Capturing and Reproducing Atmospheres. *User Interfaces & Visualization*, 27.
6. da Silveira, A. C., & Santos, C. A. S. (2022, June). Ongoing challenges of evaluating mulsemedia QoE. In *Proceedings of the 2nd Workshop on Multisensory Experiences-SensoryX'22*. SBC.

7. Ginosar, S., Rakelly, K., Sachs, S., Yin, B., & Efros, A. A. (2015). A century of portraits: A visual historical record of american high school yearbooks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* (pp. 1-7).
8. Gould, S., Fulton, R., & Koller, D. (2009, September). Decomposing a scene into geometric and semantically consistent regions. In *2009 IEEE 12th international conference on computer vision* (pp. 1-8). IEEE.
9. Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010, June). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops* (pp. 94-101). IEEE.
10. Wang, S., & Zhang, Y. (2023). Grad-CAM: understanding AI models. *Comput. Mater. Contin*, *76*(2), 1321-1324.
11. Pise, A., Vadapalli, H., & Sanders, I. (2022). Facial emotion recognition using temporal relational network: an application to E-learning. *Multimedia Tools and Applications*, *81*(19), 26633-26653
12. Llanda, C. J. R. (2019, February). Video tutoring system with automatic facial expression recognition: an enhancing approach to e-learning environment. In *Proceedings of the 2019 4th International Conference on Intelligent Information Technology* (pp. 5-9).
13. Ding, X., & Mariano, V. Y. (2024). Learning Status Recognition Method Based on Facial Expressions in e-Learning. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, *28*(4), 793-804.
14. O'shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.
15. Mayya, V., Pai, R. M., & Pai, M. M. (2016). Automatic facial expression recognition using DCNN. *Procedia Computer Science*, *93*, 453-461.
16. Clevert, D. A., Unterthiner, T., & Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*.
17. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780.
18. Staudemeyer, R. C., & Morris, E. R. (2019). Understanding LSTM--a tutorial into long short-term memory recurrent neural networks. *arXiv preprint arXiv:1909.09586*.
19. Ioffe, S., & Szegedy, C. (2015, June). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448-456). pmlr.
20. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
21. Möller, S., & Raake, A. (Eds.). (2014). Quality of experience: advanced concepts, applications, and methods. Springer.
22. Ebrahimi, T. (2009, October). Quality of multimedia experience: past, present, and future. In *Proceedings of the 17th ACM International Conference on Multimedia* (pp. 3-4).
23. Hoßfeld, T., Schatz, R., & Egger, S. (2011, September). SOS: The MOS is not enough! In *2011 third international workshop on the quality of multimedia experience* (pp. 131-136). IEEE.